chinaXiv:201804.02154v1

融合社交行为和标签行为的推荐算法研究 *

蒋　云，倪　静，房宏扬

(上海理工大学 管理学院，上海 200093)

摘　要：针对传统推荐算法忽略用户社交影响、研究角度不全面和缺乏物理解释等问题，提出一个融合社交行为和标签行为的推荐算法。首先用引力模型计算社交网络中用户节点之间的吸引力来度量用户社交行为的相似性；其次通过标签信息构建用户喜好物体模型，并使用引力公式计算喜好物体之间的引力来度量标签行为的相似性。最后，引入变量融合两方面信息，获取近邻用户，产生推荐。采用 Last。fm 数据集进行实验研究，结果说明推荐算法的准确率和召回率更高。

关键词：社交行为；标签行为；万有引力；协同过滤

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.01.0038

Study of recommended algorithms integrating social behavior and labeling behavior

Jiang Yun, Ni Jing, Fang Hongyang

(*School of Business*, *University of Shanghai for Science & Technology*, *Shanghai* 200093, *China*)

Abstract: In view of the traditional recommendation algorithm ignoring the impact of social behavior of users, the incomprehensive research perspective and lack of physical explanation, a recommendation algorithm was proposed that integrated social behavior and tagging behavior of users. Firstly, the attractiveness between user nodes in social network was calculated by gravity model to measure the similarity of users' social behavior. Secondly, the user's favorite object model was constructed by label information, the gravitation formula was also used to calculate the gravitation between favorite objects to measure the similarity of tagging behavior. Finally, the paper introduced the variables to weigh the proportion of two similar values, and then got the set of neighbors and generated recommendations. Experimental results using Last. fm dataset showed that the proposed algorithm had higher precision and recall.

Key words: social behavior; labeling behavior; gravitation; collaborative filtering

0　引言

据估计到 2020 年，全球产生的信息总数将会超 40ZB，我国的贡献率预计会占到近 21%[1]。面对数据的大规模爆发，个性化推荐技术营运而生，成功解决“信息过载”问题[2]。近几年，以 YouTube、Last.fm、微博、豆瓣等代表的社会标签系统层出不穷。成为推荐技术一个重要的研究方向。

部分学者从聚类技术[3]、模型构建[4]　、社交信任[5]等角度进行研究。还有部分学者将物理学方法应用于推荐系统中，如杨卫芳等人[6]提出一种混合热传导和物质扩散理论的方法，研究用户的活跃度，并有效改善推荐算法。王国霞在社会标签系统中通过万有引力原理分别改进用户相似度和项目相似度计量方法，分别提出用户引力[7, 8]和项目引力[9]的概念，实验获得了较其他算法更优的推荐性能。然而这些方法都忽略了用户社交行为的影响。在实际的生活中，人们的每一次选择都不可避免地会受到朋友、家人或其他信赖的人的影响。Bonhard[10]通过实验发现，用户决策时通常选择听从信任的好友的意见，而忽视系统推荐。因此，合理利用社交网络中用户社交行为将有助于提高推荐的准确度。

目前已经有一些学者在社会标签系统中融合社交行为进行推荐[11, 12]，但这些方法都缺乏一定的物理解释，且推荐的准确率还有待提高。因此，本文在综合前人研究的基础上，提出一个融合社交行为和标签行为的协同过滤推荐算法。一方面是用复杂网络理论对社交网络中的用户社交行为进行研究，利用引力原理计算基于社交行为的用户相似度；另一方面是根据用户标签行为来构建用户的兴趣向量，并采用 TF-IDF 方法计算权重，利用引力原理计算基于标签行为的用户相似度。算法的最终目的是将两方面的相似性引用变量加权求和，实现算法改进。通过本文的研究，期望能够达到提高推荐算法性能和和赋予推荐系统物理解释的目标。

收稿日期：2018-01-22；**修回日期：**2018-03-05　　**基金项目：**国家自然科学基金面上项目（71774111）

作者简介：蒋云（1993-），女，江苏泰州人，硕士研究生，主要研究方向为信息管理与决策支持系统、复杂网络（yolandaa0828@163.com）；倪静（1972-），女，副教授，博士，主要研究方向为管理信息系统、电子商务、复杂网络；房宏扬（1993-），女，硕士研究生，主要研究方向为企业供需网及其管理.

ChinaXiv合作期刊
chinaXiv:201804.02154v1

1　相关理论

1.1　传统的基于用户的协同过滤推荐

本文在传统的基于用户的协同过滤推荐基础上对用户相似性考察方式进行改进。基于用户的协同过滤推荐的流程主要包括以下三个部分：构建用户-项目评价模型、确定近邻用户、实现推荐。算法步骤如图 1 所示。

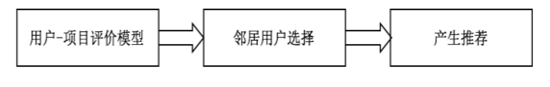


图 1　协同过滤推荐算法步骤图

1）构建用户-项目评价模型

用户对物品项目的评价有很多种表现形式。其中最为常见的是 1~5 的打分制。也可以用二值数据（0/1）来确定物品特征向量。如果用户对项目没有任何的评价行为，则对应的评价值用零值或者空值来替代。在获取评价信息的基础上，可以得到一个$n\times m$的用户-项目评价矩阵$R_{n\times m}$。其中n是用户数量，m是项目数量，R_{ij}为第i个用户U_i对第j个项目I_j的评价值。用户-项目评价矩阵是相似值考察的重要依据。

2）邻居用户选择

邻居用户的选择依赖于目标用户与系统中其他用户的相似度，相似度越大的用户对目标用户的影响越大，推荐的效果也就越好。皮尔逊相关系数、余弦相似性和修正的余弦相似性等是常用的相似度计算方法。相似度计算之后，可以采用固定值法或者是阈值法得到邻居用户的集合。

3）产生推荐

根据用户对某个项目的预测打分值来获得推荐的结果。该值计算公式如下：

$$P_{u,i}=\overline{R_u}+\frac{\sum_{v\in NB} sim(u,v)\times(R_{v,i}-\overline{R_v})}{\sum_{v\in NB}\left(sim(u,v)\right)}\qquad(1)$$

其中：$P_{u,i}$是用户 u对项目i的预测打分值，NB是待推荐用户u的近邻用户集合，$v\in NB$。$sim(u,v)$是两个用户u和v之间的相似性大小，$R_{v,i}$是用户v对项目i的打分值。变量$\overline{R_u}$和$\overline{R_v}$分别是用户u和v对系统中所有已评分项目上平均打分值。根据预测分值，按照由大到小排列，选择 Top-N 的项目物体进行推荐。结束整个推荐流程。

1.2　社会标签网络

社会标签网络的网络结构由一部图和二部图[13～15]演化而来。该网络中存在用户、项目、标签三类节点，组成三部图，如图 2 所示。

图 2　用户-项目-标签三部图

给出如上图所示的社会标签网络相关的矩阵定义为：假设用户集合为$U=(u_1,u_2,\cdots,u_n)$，n为用户总数，项目集合为$I=(i_1,i_2,\cdots,i_m)$，m为项目总数，标签集合$T=(t_1,t_2,\cdots,t_l)$，l为标签总数。三者的矩阵关系表示为：用户-项目评价矩阵R（user-item rating matrix）。如果有n个用户$U=(u_1,u_2,\cdots,u_n)$和m个项目$I=(i_1,i_2,\cdots,i_m)$，由于打分等评价行为得到一个$n\times m$的矩阵。该矩阵的行为用户，列为项目，当用户i选择了项目j并且评分为x时，矩阵中的$r_{ij}=x$，否则$r_{ij}=0$。

用户-标签频率矩阵S（user-tags frequency matrix）。假设有n个用户$U=(u_1,u_2,\cdots,u_n)$和l个标签$T=(t_1,t_2,\cdots,t_l)$，它们之间形成一个$n\times l$的矩阵。矩阵的行为用户，列为标签。当用户i使用标签g的次数为y时，矩阵中的$s_{ig}=y$，否则$s_{ig}=0$。

标签-项目频率矩阵Q（tags-item frequency matrix）。假设有l个标签$T=(t_1,t_2,\cdots,t_l)$和m个项目$I=(i_1,i_2,\cdots,i_m)$，它们之间形成一个$l\times m$的矩阵。矩阵的行为标签，列为项目。当有p个标签g标注了项目j时，矩阵中的$q_{gj}=p$，否则$q_{gj}=0$。

1.3　万有引力定律

牛顿万有引力定律认为任何两个物体在连心方向上有相互之间的吸引力。该引力的大小与它们质量的乘积成正比与它们距离的平方成反比，公式表示为

$$F=G\frac{m_1m_2}{r^2}\qquad(2)$$

其中：F表示两个物体之间的引力，G为引力常量，m_1、m_2分别表示两个物体的质量，r表示两个物体之间距离。

2　用户相似度计算

2.1　用户社交行为的相似度

互联网技术的发展和智能设备的普及，极大改变了传统的社交形式，人们可以通过网络社交来建立自己的好友关系网，形成相互之间的信任。根据在线社交网络中用户之间的社交行为可以构建用户与好友之间的矩阵关系C。假设用户集合为$U=$

ChinaXiv合作期刊

$(u_1, u_2, \cdots, u_n)$，n为用户总数，用户之间的社交好友关系形成一个$n \times n$的方阵，第i行第j列的元素c_{ij}定义如式（3）所示。

$$c_{ij} = \begin{cases} 1 & \text{用户}u_i\text{和用户}u_j\text{相互关注} \\ 0 & \text{用户}u_i\text{和用户}u_j\text{不相互关注} \end{cases} \quad (3)$$

从复杂网络的角度看，在社交网络中，用户可视为网络中的节点，用户之间的好友关系可视为网络中节点与节点之间的连边，两个关联用户共同标注过的项目物体的个数定义为边权。因此给定一个加权图 G=(V,E)，V 是节点的集合，E 是边的集合。用邻接矩阵$A = [a_{ij}]_{N \times N}$可以表示为

$$a_{ij} = \begin{cases} w_{ij}, & (v_i, v_j) \in E \\ 0, & (v_i, v_j) \notin E \end{cases} \quad (4)$$

其中：若$e_{ij} = (v_i, v_j) \in E$，$w_{ij}$表示边$e_{ij} = (v_i, v_j)$上的权值（即边权）。本文对连边的权值作出如下定义。假设两个好友用户共同标注的项目数量为b_{ij}，b_{ij}的取值存在两种情况，表示如下：

$$b_{ij} = \begin{cases} a, & I_i \cap I_j \neq \emptyset \\ 0, & I_i \cap I_i = \emptyset \end{cases} \quad (5)$$

其中：I_i表示用户u_i标注的项目集合，I_j表示用户u_j标注的项目集合。因而，定义$w_{ij} = b_{ij} + 1$，w_{ij}为大于 0 的正整数。该定义满足两个目的。一是区别于两用户在网络结构上无连接的情况；二是避免边权出现等于零的情况，为后文的计算提供保障。

将牛顿万有引力引入到复杂网络中。根据万有引力定律，用户网络中任何两个用户节点之间也存在引力作用，引力越大，说明用户之间的关系越密切，用户社交行为越相似。根据网络的结构特征，将网络中任意两个相关联的节点v_i、v_j之间的引力公式重新定义为

$$F_{ij} = G\frac{m_i m_j}{{r_{ij}}^2} \quad (6)$$

F_{ij}是节点v_i对节点v_j的引力。G为引力常量，因为本文研究的节点处于同一个网络环境中，取$G = 1$。节点的质量m_i、m_j和节点间的距离r_{ij}是计算节点间引力的关键。本文给出的定义如下。

a）质量。m_i、m_j分别表示网络中节点v_i、v_j的质量。网络中一个节点的价值首先取决于这个节点在网络中所处的位置，位置越中心的节点其价值越大。中心性反映了网络中各节点的相对重要性。因而本文采用节点的标准化度中心性来衡量网络中节点的质量，即认为一个节点的度越大，该节点越重要。因此，节点v_i的质量的计算公式为

$$m_i = k_i = \frac{\sum_{j=1}^{n} c_{ij}}{n-1} \quad (7)$$

b）距离。在加权网络中，边权按照其意义可以分为相异权和相似权。本文期望考察两个用户节点之间的相似性，因此从相似权角度出发，权值越大，两点之间的距离越小，关系也就越密切。因此边 e_{ij}的长度定义为

$$d_{ij} = \frac{1}{w_{ij}} \quad (8)$$

假设节点v_i和v_k通过两条权重分别是w_{ij}和w_{jk}的边相连，在相似权情况下节点v_i和v_k之间的距离定义为

$$d_{ik} = \frac{1}{w_{ij}} + \frac{1}{w_{jk}} \quad (9)$$

因此，根据最短路径原则，给出距离r_{ij}的数学定义如下：

假设从节点v_i到v_j，总共有p条路径。分别计算各条路径的长度分别为${d_{ij}}^1, {d_{ij}}^2, {d_{ij}}^3, \cdots, {d_{ij}}^p$，比较所有的路径，根据最短路径的原则，将$r_{ij}$定义如下：

$$r_{ij} = \min\{{d_{ij}}^1, {d_{ij}}^2, {d_{ij}}^3, \cdots, {d_{ij}}^p\} \quad (10)$$

在定义质量和距离的基础上，给出基于好友关系的相似性度量，如下：

$$\text{sim}_{(u_i,u_j)}{}^{relation} = F_{ij} = G\frac{m_i m_j}{{r_{ij}}^2} \quad (11)$$

2.2　用户标签行为的相似度

在社会标签网络中，标签可以表征用户的喜爱偏好，表达用户观点。因此从用户使用的标签的内容和频率来挖掘用户的喜好。为了计算用户标签行为的相似度，同样引入牛顿万有引力定律，给出如下的一些定义：

定义 1　项目物体。推荐系统中的项目定义为项目物体。每一个项目物体有着其质量、种类等属性。其中受用户喜欢的项目物体为用户喜好物体。

定义 2　项目微粒。

将构成物体的若干个不可分割的单元定义为项目微粒。项目微粒也具有质量、类别属性。

根据以上定义，将用户使用的标签看作该用户喜好物体的项目微粒，由这些项目微粒共同组成了用户喜好物体模型，反映用户的偏好。因此假设社会标签系统中标签集合为$T = (t_1, t_2, \cdots, t_l)$，$l$为标签总数，则对于任意一个用户$u_i$来说，用户喜好物体模型为

$$F_{u_i} = (p_{it_1}, p_{it_2}, \cdots, p_{it_l}) \quad (12)$$

其中：p_{it_l}是社会标签t_l的使用频率，表示用户u_i喜好物体中第l个项目微粒。

用户喜好物体模型一方面反映用户的喜好，另一方面被赋

chinaXiv:201804.02154v1

ChinaXiv合作期刊
chinaXiv:201804.02154v1

予物理特性。因此根据牛顿万有引力定律，用户喜好物体之间存在吸引力，该引力的大小用来衡量用户喜好物体模型的相似度，引力越大，说明用户喜好物体模型之间的相似度越大，两用户喜好物体模型中包含的项目微粒相似越多，用户标签行为越相似。因此，要计算用户喜好物体之间的引力，首先需要定义用户喜好物体的质量和它们之间的距离。

a）质量。用户喜好物体的质量由组成它们的项目微粒的质量决定。若某项目物体$item_j$含有l种项目微粒，则该项目物体的质量可以表示为一个质量向量，如下：

$$m_{item_j} = \left(m_{jt_1}, m_{jt_2}, \cdots, m_{jt_l}\right) \qquad (13)$$

其中，m_{jt_l}表示项目物体$item_j$第l个项目微粒t_l的质量，且$m_{jt_l} \geq 0$。

项目微粒的质量取决于项目微粒在用户喜好物体中的重要性程度，重要性越高，则该项目微粒的质量就越大。 那么，对于一个特定用户u_i，其喜好物体模型中某一项目微粒p_{t_l}的质量，也就是该项目微粒p_{t_l}对用户u_i的喜好物体的重要程度：

$$m\left(u_i, p_{t_l}\right) = w\left(u_i, p_{t_l}\right) \qquad (14)$$

其中，$m\left(u_i, p_{t_l}\right)$表示项目微粒$p_{t_l}$在用户$u_i$喜好物体模型中的质量，$w\left(u_i, p_{t_l}\right)$表示项目微粒$p_{t_l}$对该用户喜好物体模型的重要性参数。

重要性参数采用 TF-IDF 算法计算：

$$w\left(u_i, p_{t_l}\right) = TF_{(u_i, p_{t_l})} \times IDF_{p_{t_l}} \qquad (15)$$

其中，$TF_{(u_i, p_{t_l})}$表示项目微粒p_{t_l}在用户喜好物体中出现的频率，$IDF_{p_{t_l}}$表示，该项目微粒在所有用户的喜好物体模型中的区分能力，计算方式分别表示为：

$$TF_{(u_i, p_{t_l})} = \frac{num_{(u_i, p_{t_l})}}{num_{(u_i)}} \qquad (16)$$

其中：$num_{(u_i, p_{t_l})}$表示用户u_i使用标签t_l的次数，$num_{(u_i)}$表示用户u_i使用标签的总次数。

$$IDF_{p_{t_l}} = \log \frac{num_{user}}{num_{t_l}} \qquad (17)$$

其中：num_{user}表示推荐系统中的用户总数，num_{t_l}表示使用过标签t_l的用户的数量。

b）距离。对于任意两个用户u_i和u_j，其喜好物体之间的距离就是其对应的用户—标签频率矩阵 S 中两用户—标签向量$s_{u_i} = (s_{i1}, s_{i2}, \cdots, s_{il})$和$s_{u_j} = (s_{j1}, s_{j2}, \cdots, s_{jl})$的之间的距离，采用欧几里德距离来进行计算，数学表示如下：

$$d_{ij} = \sqrt{\sum_{t=1}^{l} (s_{it} - s_{jt})^2} \qquad (18)$$

其中：d_{ij}为用户u_i和u_j喜好物体模型之间的物理距离，s_{it}表示用户u_i使用标签 t的频率，s_{jt}表示用户u_j使用标签 t的频率。

在定义用户喜好物体模型的质量和喜好物体之间的距离的基础上，计算两个用户u_i和u_j的喜好物体之间的吸引力，从而给出基于用户喜好的相似度表达式为

$$sim_{(u_i,u_j)}{}^{preference} F_{ij}{}' = G \frac{m_{item_i} \cdot m_{item_j}}{d_{ij}{}^2} \qquad (19)$$

其中：$F_{ij}{}'$表示两个用户喜好物体模型的吸引力。G为引力常量，本研究默认为常量 1。m_{item_i}和m_{item_j}分别表示两个用户的喜好物体的质量，d_{ij}为用户u_i和u_j喜好物体模型之间的物理距离。

2.3　融合社交行为和标签行为的用户相似度

综合前文研究，提出融合社交行为和标签行为的用户相似度计算方法如下：

$$sim_{(u_i,u_j)} = \alpha sim_{(u_i,u_j)}{}^{relation} + (1-\alpha) sim_{(u_i,u_j)}{}^{preference} \qquad (20)$$

其中：$0 \leq \alpha \leq 1$，用于权衡社交行为和标签行为对用户相似值的影响程度。α数值具有不确定性，将视社会标签网络的具体情况而定。

3　融合社交行为和标签行为的推荐

本文在按照融合社交行为和标签行为的相似值考量方法求得用户相似值之后，使用 Top-*K* 法获取目标用户的近邻集合 NB。采用如下的方法产生推荐的结果：

首先假设待推荐用户u的前 Top-K 邻居组成的集合为$NB = \{v_1, v_2, \cdots, v_k\}$，对于任一$v_i \epsilon$NB，获取其有过评价的项目物体集合$Item_i$，依次遍历所有邻居用户，将各个邻居用户评价过的项目集合$Item_i(i = 1,2, \cdots, k)$组成新的集合 Item-Group，且保留重复项目，然后对 Item-Group 中的项目依据项目类别进行 count 计数，根据阈值法的基本思想，筛选出个数超过两个的项目，即$\text{count} \geq 2$项目组成新的集合 Item-Recommend 推荐给用户。

4　实验结果与分析

4.1　实验数据

本文使用的实验数据是 Last.fm 数据集。该数据集包含 1892 个用户，17632 位歌手和 11946 个标签，共产生 12712 条朋友关系记录和 184679 个标签行为记录。用户好友社交关系网络如图 3 所示。总共构成 20 个连通子图，其中最大的连通子图共

ChinaXiv合作期刊

chinaXiv:201804.02154v1

包含 1843 个用户，占了整个数据集的 97.4%。

实验评估前，首先针对数据集中存在噪音数据现象，本文筛选出部分数据作为实验数据：针对 Last.fm 数据集，要求每一个用户至少给 20 个音乐家标注过标签。经过筛选，最后保留 619 个用户及其所有的好友信息记录和标签行为记录。其次，将这组数据划分为两部分：训练集和测试集。本文使用交叉验证法总共进行五次实验，得到 5 组实验结果，然后求平均值，得到最后的结果。

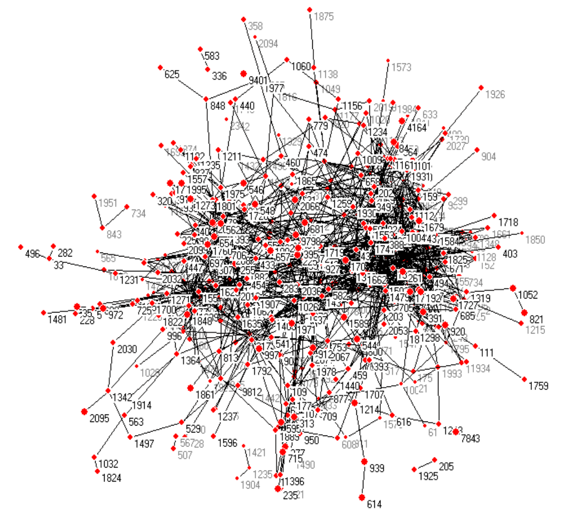


图 3　Last.fm 数据集用户社交关系网络图

4.2　实验评估指标

本文使用准确率 (precision)和召回率（recall）来考察推荐模型的质量和效果。准确率表征的是项目物品被成功推荐的比例。计算公式如下：

$$precision = \frac{n_{like}}{n_{recommend}} \qquad (21)$$

其中：$n_{recommend}$表示给推荐项目集合的元素个数，n_{like}表示在推荐的项目中，受该用户喜欢的项目总数。准确率的值越大，模型效果越好。

召回率的计算公式为

$$recall = \frac{n_{like}}{n_{test}} \qquad (22)$$

其中，n_{test}表示测试集中项目物体的数量，n_{like}表示在推荐的项目物体中，用户实际感兴趣的项目物体的数量。召回率的值越大，模型效果越好。

4.3　实验结果分析

本文的推荐算法中权重值α和用户邻居用户 K 值都是不确定的，因此，本实验首先对α和 K 的合理取值进行实验验证，取得最优的实验效果，然后在α值取得最优的情况下，将本文提出的算法与其他推荐算法进行性能比较，得出结论。

1）权重值α对算法的影响

为了权衡α对算法的影响，分别将α的取值设为 0.0，0.1，0.2，0.3，…，1.0 进行实验，并且为了消除邻居用户数量对推荐结果的影响，实验过程中依次将邻居用户的数量设为 2,3,4, …,10，共进行了 9 组实验。实验结果如下图 4。可以看出不同的α值对本文融合社交行为和标签行为的推荐算法（SNUB-CF 算法）准确率的影响。

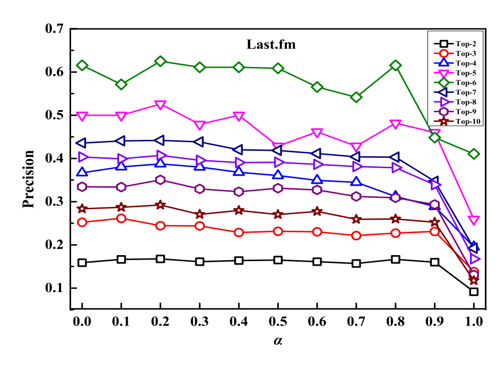


图 4　不同α对 SNUB-CF 算法的准确率的影响

从图 4 中可以看出，不管邻居用户 K 的取值，随着α从 0.0 开始不断增加，SNUB-CF 算法准确率波动的趋势总体上是一致的，呈现先上升，达到最高点，再缓慢下降的趋势。当α取值是 0.2 时，推荐模型的准确率的值最好。也就是说在本文给出的推荐算法的模型中，社交行为与标签行为对最后推荐结果的贡献比例是 1:4。说明用户在进行项目选择时，信任好友对其的影响不大，用户更多的还是考虑到自身的喜好。当α在 0.8～1.0 的范围内变化时，推荐准确率急剧下降。尤其是当$\alpha = 1.0$时，即只考虑用户社交行为时，与$\alpha = 0.9$时相比，推荐的准确率下降 40～50%，这进一步说明如果不考虑用户标签行为，单纯考虑社交行为时，推荐的准确率将大大下降。

2）Top-K 取值对算法的影响

根据本文产生推荐的方法，将 K 的取值设为 2,3,4, …,10，实验观察推荐准确率随着 K 值不断变化的规律。实验结果如图 5。结果表明，不管权重值α的取值，随着 K 值的不断增加，SNUB-CF 算法准确率呈现先上升，达到最高点，再下降的趋势。

在本文实验环境下，当 K 的值在 2～5 区间内浮动时，算法准确率不断上升。当 K=6 时，准确率的值最高。当 K 的值在 6～10 区间内浮动时，推荐的准确率又开始下降。因此，实验说明邻居数量的大小能够影响推荐的质量，但并不是邻居数量越多，推荐的质量越好。对于本实验的数据环境而言，当目标

chinaXiv:201804.02154v1

用户的近邻取值为 6 的时候，实验结果最好。

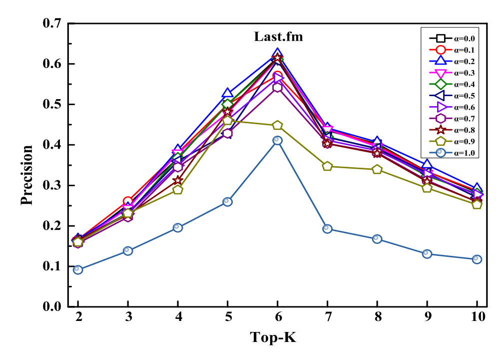


图 5 不同的 K 值情况下 SNUB-CF 算法的准确率

对于 K 值较大时，准确率降低的问题，可对本文提出的产生推荐的过程进行改进。因为当邻居用户比较多的时候，生成的 Item-Group 的集合中项目数量就会增多，而此时该集合中 $count \geq 2$的项目数量也就较多，但是随着近邻用户数量的增加会导致与目标用户相似值较小的用户也包含在其中，从而导致 Item-Recommend 的集合中推荐的项目不一定是目标用户喜欢的，因此在推荐项目物体数量较多且被正确推荐项目物体数量较小的情况下，结果的准确率大大降低。为了改善这一问题，可以对 count 进行阈值调整，提高项目的筛选要求。

当 K=8 时，取$count \geq 2$和$count \geq 3$分别计算推荐的准确率，结果如下图 6。从该图中可以看出，当 count 的阈值提高时，进一步限制了推荐项目的范围，推荐的准确率也相应提高。当 K=9,10,11,…时，实验也呈现出此规律。因此，能够采用调整推荐项目物体的阈值来提升推荐结果的质量。

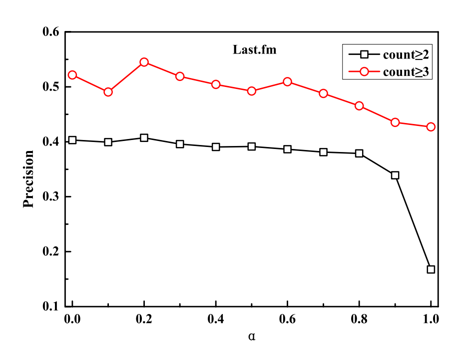


图 6　K=8 时不同 count 值情况下 SNUB-CF 算法的准确率

3）不同算法的推荐性能比较

为了验证本文提出推荐算法的有效性，将文 SNUB-CF 算法与其他算法进行对比分析，分别是：FT-CF 算法[12]（Hybrid Collaborative Filtering Recommendation Algorithm Based on Friendships and Tag，基于好友关系和标签的混合协同过滤算法）、PRT-CF 算法[16]（personalized resource recommendation based on tags and collaborative filtering recommendation，基于标签和协同过滤的个性化资源推荐）、UGBCF 算法[8]（collaborative filtering recommendation algorithm based on user's gravitation，基于用户引力的协同过滤推荐算法）、Social-CF 算法代表本文只考虑用户社交行为的推荐算法、Tag-CF 算法代表本文只考虑用户标签行为的推荐算法。通过实验，得到以上方法的准确率值和召回率的值，分别如图 7 和 8 所示。

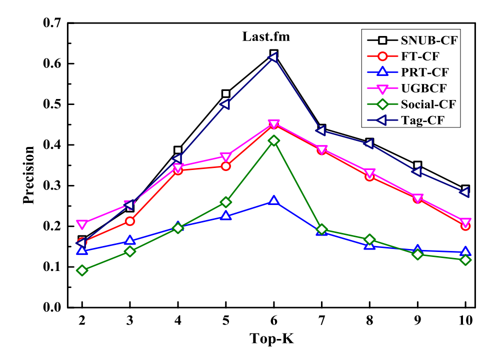


图 7 不同 K 值情况下各算法的准确率

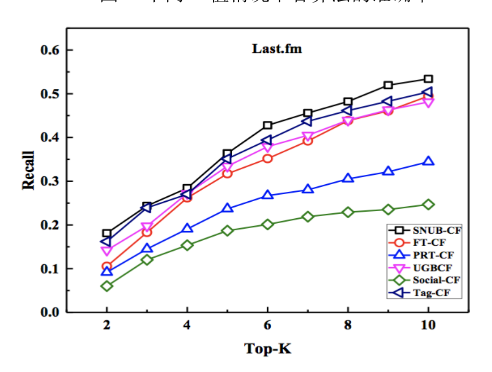


图 8 不同 K 值情况下各算法的召回率

如图 7 和 8 呈现的内容，随着邻居用户 K 值的不断增加，各个算法在指标上的变化规律趋于一致。对于准确率，随着 K 值的增加，其呈现先增大后减小的趋势，进一步验证前文的 Top K=6 时实验效果最好的结论；对于召回率，随着 K 值的增加，召回率呈现不断增长的趋势，但是增长的幅度又快到慢，逐渐趋于平缓。

在不同的算法之间，如图 7 和 8 所示，本文的 SNUB-CF 算法在准确率和召回率方面都要优于其他算法。一方面，与 FT-CF 算法对比，虽然 FT-CF 算法也从好友关系和标签两个角度综合考虑用户之间的相似度，但是该算法的准确率和召回率均低于 SNUB-CF 算法，说明本文使用万有引力模型考察用户相似值的方法更加精确。另一方面，与 PRT-CF 算法、UGBCF 算法和 Tag-CF 算法对比，这三个算法都仅仅从用户使用的标签信息出发，挖掘用户的兴趣偏好，且 UGBCF 算法同样使用引力模型计算用户基于标签信息的相似度，但是试验结果表明，本文的 SNUB-CF 算法的准确率和召回率都优于这三个算法，

ChinaXiv合作期刊
chinaXiv:201804.02154v1

说明在社会标签网络中，仅仅考虑用户的兴趣偏好具有一定的片面性，如果同时考虑用户之间的社交行为，那么推荐的效果将更为优越。

5　结束语

从实验结果看，SNUB-CF 算法具有良好的推荐性能，说明在豆瓣、Last.fm 等带有社交性质的社会标签网络中，融合用户社交行为和标签行为能够更加全面地考察用户之间的相似度。且本文创新性地将牛顿万有引力定律与复杂网络相结合，提出社交网络中用户引力，并赋予推荐系统物理解释。实验表明，SNUB-CF 算法在计算用户相似值时，考虑角度更为全面，计算方法更为精确，因而获得的邻居用户更为相似，推荐的准确率和召回率更高。

文章主要讨论静态网络的协同过滤推荐算法，然而伴随着系统中用户、项目和标签的数量持续增多，数据稀疏问题等会导致推荐算法的性能降低。结合时序网络的特征，分析总结出在不同的时间切片上用户的相似性特点，从而提出在动态网络中有效的推荐算法模型文章进一步研究的重点。

此外，本文仅仅是对标签信息进定量分析，而标签数据中会包含丰富的语义信息。如果能够借助自然语言处理的语义分析模型及语义分析工具进一步挖掘标签中的信息，将能更好地提高推荐的质量，这也是本文的下一步研究方向。

参考文献：

[1] 上海证券报. 全球信息数据量逐年猛增 IDC 产业迎来发展新机遇 [EB/OL]. (2016-08-05) [2017-11-12]. http://stock. hexun. com/2016-08-05/185341978. html.

[2] 王国霞, 刘贺平. 个性化推荐系统综述 [J]. 计算机工程与应用, 2012, 48 (7): 66-76.

[3] 刘如娟. 基于标签聚类与用户模型的个性化推荐方法研究 [J]. 现代情报, 2016, 36 (6): 74-78.

[4] 张彬彬, 林丕源, 黄沛杰. 基于 LDA 的社会化标签系统推荐技术 [J]. 计算机工程与设计, 2016, 37 (10): 2722-2727.

[5] 杨亚东, 熊庆国. 基于动态标签偏好信任概率矩阵分解模型的推荐算法 [J]. 计算机工程, 2017, 43 (10): 160-166.

[6] 杨卫芳, 李学明, 乔保学. 改进的热传导和物质扩散混合推荐算法 [J]. 计算机工程, 2017, 43 (3): 247-252.

[7] 王国霞. 基于万有引力和随机行走的推荐算法研究 [J]. 计算机应用研究, 2016, 33 (8): 2278-2281.

[8] 王国霞. 基于用户引力的协同过滤推荐算法 [J]. 计算机应用研究, 2016, 33 (11): 3329-3333.

[9] 王国霞, 刘贺平, 李擎. 基于万有引力的个性化推荐算法 [J]. 北京科技大学学报, 2015 (2): 255-259.

[10] Bonhard P, Sasse M A. 'Knowing me, knowing you'— Using profiles and social networking to improve recommender systems [J]. Bt Technology Journal, 2006, 24 (3): 84-98.

[11] 丁小焕, 彭甫镕, 王琼, 等. 融合朋友关系和标签信息的张量分解推荐算法 [J]. 计算机应用, 2015, 35 (7): 1979-1983.

[12] 曾安, 徐小强. 基于好友关系和标签的混合协同过滤算法 [J]. 计算机科学, 2017, 44 (8): 246-251.

[13] Zhou T, Kuscsik Z, Liu J G, *et al.* Solving the apparent diversity-accuracy dilemma of recommender systems [J]. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107 (10): 4511.

[14] Liu J G, Zhou T, Che H A, *et al.* Effects of high-order correlations on personalized recommendations for bipartite networks [J]. Physica A Statistical Mechanics & Its Applications, 2010, 389 (4): 881-886.

[15] Shang M, Lu L, Zhang Y C, *et al.* Empirical analysis of web-based user-object bipartite networks [J]. Europhysics Letters, 2012, 90 (4): 1303-1324.

[16] 蔡强, 韩东梅, 李海生, 等. 基于标签和协同过滤的个性化资源推荐 [J]. 计算机科学, 2014, 41 (1): 69-71.